ED 408 336 TM 026 589

AUTHOR	Daniel, Larry G.
TITLE	Statistical Significance Testing in "Educational and Psychological Measurement" and Other Journals.
PUB DATE	Mar 97
NOTE	33p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, March 25-27, 1997).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Editing; *Educational Assessment; Policy; Research Problems; *Scholarly Journals; *Social Science Research; *Statistical Significance; *Test Use
IDENTIFIERS	*Educational and Psychological Measurement

ABSTRACT

Statistical significance tests (SSTs) have been the object of much controversy among social scientists. Proponents have hailed SSTs as an objective means for minimizing the likelihood that chance factors have contributed to research results. Critics have both questioned the logic underlying SSTs and bemoaned the widespread misapplication and misinterpretation of the results of these tests. This paper offers a framework for remedying some of the common problems associated with SSTs via modification of journal editorial policies. The controversy surrounding SSTs is reviewed, with attention given to both historical and more contemporary criticisms of bad practices associated with misuse of SSTs. Examples from the editorial policies of "Educational and Psychological Measurement" and several other journals that have established guidelines for reporting results of SSTs are discussed, and suggestions are provided regarding additional ways that educational journals may address the problem. These guidelines focus on selecting qualified editors and reviewers, defining policies about use of SSTs that are in line with those of the American Psychological Association, and stressing effect size reporting. An appendix presents a manuscript review form. (Contains 61 references.) (Author/SLD)

stsigepm.aer

Running Head: STATISTICAL SIGNIFICANCE TESTING

Statistical Significance Testing in *Educational and Psychological Measurement* and Other Journals

Larry G. Daniel
University of Southern Mississippi

Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL, March 24-28, 1997.

ABSTRACT

Statistical significance tests (SSTs) have been the object of much controversy among social scientists. Proponents have hailed SSTs as an objective means for minimizing the likelihood that chance factors have contributed to research results; critics have both questioned the logic underlying SSTs and bemoaned the widespread misapplication and misinterpretation of the results of these tests. The present paper offers a framework for remedying some of the common problems associated with SSTs via modification of journal editorial policies. The controversy surrounding SSTs is overviewed, with attention given to both historical and more contemporary criticisms of bad practices associated with misuse of SSTs. Examples from the editorial policies of Educational and Psychological Measurement and several other journals that have established guidelines for reporting results of SSTs are overviewed, and suggestions are provided regarding additional ways that educational journals may address the problem.

Statistical Significance Testing in *Educational and Psychological Measurement* and Other Journals

Statistical significance testing has existed in some form for approximately 300 years (Huberty, 1993), and has served an important purpose in the advancement of inquiry in the social sciences. However, there has been much controversy over the misuse and misinterpretation of statistical significance testing. Pedhazur and Schmelkin (1991, p. 198) noted, "Probably few methodological issues have generated as much controversy among sociobehavioral scientists as the use of [statistical significance] tests." This controversy has been evident in social science literature for some time, and many of the articles and books exposing the problems with statistical significance have aroused remarkable interest within the field. In fact, at least two articles on the topic appeared in a list of works rated by the editorial board members of *Educational and Psychological Measurement* as most influential to the field of social science measurement (Thompson & Daniel, 1996). Interestingly, the criticisms of statistical significance testing have been pronounced to the point that, when one reviews the literature, "it is more difficult to find specific arguments for significance tests than it is to find arguments decrying their use" (Henkel, 1976, p. 87).

Thompson (1987b) noted that researchers are increasingly becoming aware of the problem of over-reliance on statistical significance tests (referred to herein as "SSTs"). However, despite the influence of the many works critical of practices associated with SSTs, many of the problems raised by the critics are still prevalent. Researchers have inappropriately utilized statistical significance as a means for illustrating the importance of their findings and have attributed to statistical significance testing qualities it does not possess. Reflecting on this problem, one psychological researcher observed, "the test of significance does not provide the information concerning psychological phenomena characteristically attributed to it; . . .a great deal of mischief has been associated with its use (Bakan, 1966, p. 423).

Because SSTs have been so frequently misapplied, some reflective researchers (e.g., Carver, 1978; Cronbach, 1975; Meehl, 1978; Shulman, 1970) have recommended that SSTs be completely abandoned as a method for evaluating statistical results. In fact Carver (1983) not only recommended abandoning statistical significance testing, but referred to it as a "corrupt form of the scientific method" (p. 288). Interestingly, the American Psychological Association has now appointed its Task Force on Statistical Affairs, which will consider among other actions recommending less or even no use of statistical significance testing within APA journals (Shea, 1996). On the other hand, SSTs still have support from a number of reflective researchers who acknowledge their limitations, but also see the value of the tests when appropriately applied. For example, Mohr

(1990, p. 74) reasoned, "one cannot be a slave to significance tests. But as a first approximation to what is going on in a mass of data, it is difficult to beat this particular metric for communication and versatility." In similar fashion, Huberty (1987, p. 7) maintained, "there is nothing wrong with statistical tests themselves! When used as guides and indicators, as opposed to a means of arriving at definitive answers, they are okay."

"Statistical Significance" Versus "Importance"

A major controversy in the interpretation of SSTs has been "the ingenuous assumption that a statistically significant result is necessarily a noteworthy result (Daniel, 1997, p. 106). Thoughtful social scientists (e.g., Chow, 1988; Gold, 1969; Winch & Campbell, 1969; Shaver, 1993) have long recognized this problem. For example, as early as 1931, Tyler had already begun to note a trend toward the misinterpretation of statistical significance:

> The interpretations which have commonly been drawn from recent studies indicate clearly that we are prone to conceive of statistical significance as equivalent to social significance. These two terms are essentially different and ought not to be confused. . . . Differences which are statistically significant are not always socially important. The corollary is also true: differences which are not shown to be statistically significant may nevertheless be socially significant. (pp. 115-117)

A decade later, Berkson (1942, p. 325) remarked, "statistics, as it is taught at present in the dominant school, consists almost entirely of tests of significance." Likewise, by 1951, Yates observed, ". . .scientific workers have often regarded the execution of a test of significance on an experiment as the ultimate objective. Results are significant or not significant and this is the end of it" (p. 33). Similarly, Kish (1959) bemoaned the fact that too much of the research he had seen was presented "at the primitive level" (p. 338). Twenty years later, Kerlinger (1979, pp. 318-319) recognized that the problem still existed:

> . . .statistical significance says little or nothing about the magnitude of a difference or of a relation. With a large number of subjects. . .tests of significance show statistical significance even when a difference between means is quite small, perhaps trivial, or a correlation coefficient is very small and trivial. . . . To use statistics adequately, one must understand the principles involved and be able to judge whether obtained results are statistically significant <u>and</u> whether they are meaningful in the particular research context. (emphasis in original)

Contemporary scholars continue to recognize the existence of this problem. For instance, Thompson (1996) and Pedhazur and Schmelkin (1991) credit the continuance of the misperception, in part, to the tendency of researchers to utilize and journals to publish manuscripts containing the term "significant" rather than "statistically significant"; thus, it becomes "common practice to drop the word 'statistical,' and speak instead of 'significant differences,' 'significant correlations,' and the like" (Pedhazur & Schmelkin, 1991, p. 202). Similarly, Schafer (1993) noted, "I hope most researchers understand that *significant* (statistically) and important are two different things. Surely the term *significant* was ill chosen" (p. 387--emphasis in original).

SSTs and Sample Size

Most tests of statistical significance utilize some test statistic (e.g., *F*, *t*, chi-square) with a known distribution. A statistical significance test is simply a comparison of the value for a particular test statistic based on results of a given analysis with the values that are "typical" for the given test statistic. The computational methods utilized in generating these test statistics yield larger values as sample size is increased. In other words, a large sample is more likely to guarantee the researcher a statistically significant result than a small sample is.

For example, a researcher might conduct an educational experiment in which students are randomly assigned to two different instructional settings and are then evaluated on an outcome achievement measure. This researcher might utilize an analysis of variance test to evaluate the result of the experiment. Prior to conducting the test (and the experiment), the researcher would propose a null hypothesis of no difference between persons in varied experimental conditions and then compute an *F* statistic by which the null hypothesis may be evaluated. *F* is an intuitively-simple ratio statistic based on the quotient of the mean square for the effect(s) divided by the mean square for the error term. Since mean squares are the result of dividing the sum of squares for each effect by its degrees of freedom, the mean square for the error term will get smaller as the sample size is increased and will, in turn, serve as a smaller divisor for the mean square for the effect, yielding a larger value for the *F* statistic. In the present example (a two-group, one-way ANOVA) a sample of 3,002 would be five times as likely to yield a statistically significant result as a sample of 602 simply due to a larger number of error degrees of freedom (3,000 versus 600). In fact, with a sample as large as 3,002, even inordinately trivial differences between the two groups would be statistically significant. Large *F* values are less likely to have occurred by chance; therefore, the *p* value (likelihood of a chance result) associated with a large *F* will be small.

As this example illustrates, an SST is largely a test of whether or not the sample is large, a fact that the researcher knows even before the experiment takes place. Put simply, "Statistical significance testing can involve a tautological logic in which tired researchers, having collected data from hundreds of subjects, then conduct a statistical test to evaluate whether there were a lot of subjects. . ." (Thompson, 1992, p.

436). Some 60 years ago, Berkson (1938, pp. 526-527) exposed this circuitous logic based on his own observation of statistical significance values associated with chi-square tests with approximately 200,000 subjects:

> . . .an observant statistician who has had any considerable experience with applying the chi-square test repeatedly will agree with my statement that, as a matter of observation, when the numbers in the data are quite large, the *P*'s tend to come out small. . . and no matter how small the discrepancy between the normal curve and the true curve of observations, the chi-square *P* will be small if the sample has a sufficiently large number of observations it. . . . If, then, we know in advance the *P* that will result from an application of a chi-square test to a large sample, there would seem to be no use in doing it on a smaller one. But since the result of the former test is known, it is no test at all!

Misinterpretation of the Meaning of "Statistically Significant"

An analysis of past and current social science literature will yield evidence of at least five common misperceptions about the meaning of "statistically significant." The first of these, that "statistically significant" means "important," has already been addressed herein. Four additional misperceptions will also be discussed briefly: (a) the misperception that statistical significance informs the researcher as to the likelihood that a given result will be replicable ("the replicability fantasy"--

Carver, 1978); (b) the misperception that statistical significance informs the researcher as to the likelihood that results were due to chance (or, as Carver [1978, p. 383] termed it, "the odds-against-chance fantasy"); (c) the misperception that a statistically significant result indicates the likelihood that the sample employed is representative of the population; and (d) the misperception that statistical significance is the best way to evaluate statistical results.

<u>SSTs and replicability</u>. Despite misperceptions to the contrary, the logic of statistical significance testing is NOT a means for assessing result replicability (Carver, 1978; Thompson, 1993). Statistical significance simply indicates the probability that the null hypothesis is true in the population. However, Thompson (1993) provides discussion of procedures that may provide an estimate of replicability. These procedures (cross validation, jackknife methods, and bootstrap methods) all involve sample splitting logics and allow for the computation of statistical estimators across multiple configurations of the same sample in a single study. Even though these methods are biased to some degree (a single sample is utilized in each procedures), they represent the next best alternative to conducting a replication of the given study (Daniel, 1992b). Ferrell (1992) demonstrated how results from a single multiple regression analysis can be cross validated by randomly splitting the original sample and predicting dependent variable scores for each half of the sample using the opposite group's weights. Daniel

(1989b) and Tucker and Daniel (1992) used a similar logic in their analyses of the generalizability of results with the sophisticated "jackknife" procedure. Similar heuristic presentations of the computer-intensive "bootstrap" logic are also available in the extant literature (e.g., Daniel, 1992b).

SSTs and odds against chance. This common misperception is based on the naive perception that statistical significance measures the degree to which results occur by chance. According to this erroneous belief, a result that is statistically significant at the .01 level would be only 1% likely to have occurred by chance. This fallacy was exposed by Carver (1978, p. 383):

> . . .the *p* value is the probability of getting the research results when it is first assumed that it is actually true that chance caused the results. It is therefore impossible for the *p* value to be the probability that chance caused the mean difference between two research groups since (a) the *p* value was calculated by assuming that the probability was 1.00 that chance did cause the mean difference, and (b) the *p* value is used to decide whether to accept or reject the idea that probability is 1.00 that chance caused the mean difference.

SSTs and sampling. This misperception states that the purpose of statistical significance testing is to determine the degree to which the sample represents the population.

Representativeness of the sample cannot be evaluated with an SST; the only way to estimate if a sample is representative is to carefully select the sample. In fact, the statistical significance test is better conceptualized as answering the question, "*If* the sample represents the population, how likely is the obtained result?"

SSTs and evaluation of results. This final misperception, which states that the best (or correct) way to evaluate the statistical results is to consult the statistical significance test, often accompanies the "importance" misperception, but actually may go a step beyond the importance misperception in its corruptness. The importance misperception, as previously noted, simply places emphasis on the wrong thing. For example, the researcher might present a table of correlations, but in interpreting and discussing the results, only discuss whether or not each test yielded a statistically significant result, making momentous claims for statistically significant correlations no matter how small and ignoring statistically nonsignificant values no matter how large. In this case, the knowledgeable reader could still look at the correlations and draw more appropriate conclusions based on the magnitude of the *r* values. However, if the researcher were motivated by the "result evaluation" misperception, he or she might go so far as to fail to report the actual correlation values, stating only that certain relationships were statistically significant. Likewise, in the case of an analysis of variance, this researcher might simply report the *F* statistic and its *p* value. Thompson (1994) discusses several suggestions for improvement of these practices, including the reporting of (a) effect sizes for all parametric analyses, (b) "what if" analyses "indicating at what different sample size a given result would become statistically significant or would have no longer been statistically" (p. 845). In regard to (b), Morse (1991) has designed a PC-compatible computer program for assessing the sensitivity of results to sample size. Moreover, in the cases in which statistically nonsignificant results are obtained, researchers should consider conducting a statistical power analyses (Cohen, 1988).

Journal Policies and Statistical Significance

As most educational researchers are aware, social science journals have for years had a bias towards accepting manuscripts documenting statistically significant findings and rejecting those with statistically nonsignificant findings. One editor even went so far as to boast that he had made it a practice to avoid accepting for publication results that were statistically significant at the .05 level, desiring instead that results reached at least the .01 level (Melton, 1962). Because of this editorial bias, many researchers (e.g., Mahoney, 1976) have paid homage to SSTs in public while realizing their limitations in private. As one observer noted a generation ago, "Too, often. . .even wise and ingenious investigators, for varieties of reasons, not the least of which are the editorial policies of our major psychological journals, . . .tend to credit the test of significance with properties it does not have" (Bakan, 1966, p. 423).

According to many researchers (e.g., Neuliep, 1991; Shaver, 1993), this bias against studies that do not report statistical significance or that present results that did not meet the critical alpha level still exists. Shaver (1993, p. 310) eloquently summarized this problem:

> Publication is crucial to success in the academic world. Researchers shape their studies, as well as the manuscripts reporting the research, according to accepted ways of thinking about analysis and interpretation and to fit their perceptions of what is publishable. To break from the mold might be courageous, but, at least for the untenured faculty member with some commitment to self-interest, foolish.

Because this bias is so prevalent, it is not uncommon to find examples in the literature of studies that report results that are statistically nonsignificant with the disclaimer that the results "approached significance." Thompson (1993a) reported a somewhat humorous, though poignant, response by one journal editor to this type of statement: "How do you know your results were not working very hard to *avoid* being statistically significant?" (p. 285--emphasis in original).

Likewise, results that are statistically significant at a conservative alpha level (e.g, .001), are with some frequency referred to as "highly significant," perhaps with the authors' intent being to make a more favorable impression on some journal editors and readers than they could make by simply saying that the result was statistically significant, period. This practice, along with the even more widespread affinity for placing more and more zeroes to the right of the decimal in an attempt to make a calculated p appear more noteworthy, has absolutely nothing to do with the practical significance of the result. The latter practice has often been the focus of tongue-in-cheek comments. For example, Popham (1993, p. 266) noted, "Some evaluators report their probabilities so that they look like the scoreboard for a no-hit baseball game (e.g., p < .000000001)"; Campbell (1982, p. 698) quipped, "It is almost impossible to drag authors away from their *p* values, and the more zeroes after the decimal point, the harder people cling to them"; and McDonald (1985, p. 20), referring to the tendency of authors to place varying numbers of stars after statistical results reported in tabular form as a means for displaying differing levels of statistical significance, bantered that the practice resembled "grading of hotels in guidebooks."

If improvements are to be made in the interpretation and use of SSTs, professional journals (Rozeboom, 1960), and, more particularly, their editors will no doubt have to assume a leadership role in the effort. As Shaver (1993) articulated it, "As gatekeepers to the publishing realm, journal editors have tremendous power. . .[and perhaps should] become crusaders for an agnostic, if not atheistic, approach to tests of statistical significance" (pp. 310-311). Hence, Carver (1978, 1993) and Kupfersmid (1988) suggested that journal editors are the most likely candidates to promote an end to the misuse and misinterpretation of SSTs.

Considering this, it is encouraging to note that at least some journals have begun to adopt policies relative to statistical significance testing that address some of the problems discussed here. For several years, Measurement and Evaluation in Counseling and Development (1992, p. 143) has included three specific (and appropriate) author guidelines related to statistical significance testing:

> 8. Authors are encouraged to assist readers in interpreting statistical significance of their results. For example, results many be indexed to sample size. An author may wish to say, "this correlation coefficient would have still been statistically significant even if the sample had been as small as *n* = 33," or "this correlation coefficient would have been statistically significant if sample size had been as small as *n* = 138."
>
> 9. Authors are encouraged to provide readers with effect size estimates as well as statistical significance tests. For example, in an analysis of variance authors may wish to report eta squared or omega squared. Standardized effect size estimates (the difference between the intervention group mean minus control group mean

divided by the control group standard deviation) are also helpful in interpretation.

10. Studies in which statistical significance is not achieved will still be seriously considered for publication if power estimates of protection against Type II error are reported and reasonable protection is available.

Educational and Psychological Measurement (EPM) has developed a similar set of editorial policies (Thompson, 1994) which are presently in their third year of implementation. These guidelines do not for the most part ban the use of SSTs from being included in author's manuscripts, but rather request that authors report other information along with the SST results. Specifically, these editorial guidelines include the following:

1. Requirement that authors use "statistically significant" and not merely "significant" in discussing results.
2. Requirement that tests of statistical significance NOT accompany validity and reliability coefficients (Daniel & Witta, 1997; Huck & Cormier, 1996). This is the one scenario in which SSTs are expressly forbidden according to *EPM* editorial policy.
3. Requirement that all statistical significance tests be accompanied by effect size estimates.
4. Suggestion that authors may wish to report the "what if" analyses alluded to earlier. These analyses should

indicate "at what different sample size a given fixed effect would become statistically significant or would have no longer been statistically significant" (Thompson, 1994, p. 845).

5. Suggestion that authors report external replicability analyses via use of data from multiple samples or else internal replicability analyses via use of cross-validation, jackknife, or bootstrap procedures.

A number of efforts have been utilized by the *EPM* editors to help both authors and reviewers become familiar with the guidelines. For the first two years that these guidelines were in force, copies of the guidelines editorial (Thompson, 1994) were sent to every author along with the manuscript acceptance letter. Also, the current manuscript acknowledgement letter includes a reference to this and two other author guidelines editorials the journal has published (Thompson, 1995; Thompson & Daniel, 1996), and it directs the author to refer to the several editorials to determine if their manuscripts meet editorial policy. More recently, the several editorials have been made available via the Internet at Web address: "http://acs.tamu.edu/bbt6147/".

In addition to this widescale distribution policy, the guidelines are referenced on each review form (see Appendix A) sent to the masked reviewers. As a part of the review process, reviewers must determine if manuscripts contain material that is in violation of the editorial policies relative to statistical significance testing and several other methodological issues. To assure that reviewers will take this responsibility seriously, several questions relative to the guidelines editorials are included on the review form and must be answered by the reviewers. No manuscripts are accepted for publication by either of the two current editors if they violate these policies, although these violations do not necessarily call for outright rejection of the manuscripts. It is the hope of the editors that this comprehensive policy will over time make a serious impact on *EPM* authors' and readers' ideas about correct practice in reporting the results of SSTs.

Recommendations for Journal Editors

As the previous discussion has illustrated, there is a clear trend among social science journal editors to either reject or demand revision of manuscripts in which authors employ loose language relative to their interpretations of SSTs or else overinterpret the results of these tests. Pursuant to the continuance of this trend, the following 10 recommendations are offered to journal editors and scholars at large as a means for encouraging better practices in educational journals and other social science journals.

1. Implement editor and reviewer selection policies. First, following the suggestions of Carver (1978, 1993) and Shaver (1993), it would be wise for professional associations and publishers who hire/appoint editors for their publications to require potential editors to

submit statements relative to their positions on statistical significance testing. Journal editors might also require a similar statement from persons who are being considered as members of editorial review boards.

2. Develop guidelines governing SSTs. Each editor should adopt a set of editorial guidelines that will promote correct practice relative to the use of SSTs. The *Measurement and Evaluation in Counseling and Development* and *Educational and Psychological Measurement* guidelines referenced in this paper could serve as a model for policies developed for other journals.

3. Develop a means for making the policies known to all involved. Editors should implement a mechanism whereby authors and reviewers will be likely to remember and reflect upon the policies. The procedures mentioned previously that are currently utilized by the editors of *Educational and Psychological Measurement* might serve as a model that could be adapted to the needs of a given journal.

4. Enforce current APA guidelines for reporting SSTs. Considering that most journals in education and psychology utilize APA publication guidelines, editors could simply make it a requirement that the guidelines for reporting results of SSTs included in the fourth

edition *Publication Manual of the American Psychological Association* (APA, 1994, pp. 17-18) be followed. Although the third edition *Publication Manual* was criticized for using statistical significance reporting examples that were flawed (Pedhazur & Schmelkin, 1991; Shaver, 1993), the fourth edition includes appropriate examples as well as suggestions encouraging authors to also report effect size estimates.

5. <u>Require authors to use "statistically" before "significant."</u> Despite the fact that some journal editors will be resistant to the suggestion (see, for example, Levin, 1993), requiring authors to routinely use the term "statistically significant" rather than simply "significant" (cf. Carver, 1993; Cohen, 1990; Daniel, 1988; Shaver, 1993) when referring to research findings will do much to avoid the "statistical significance as importance" problem, and to make it clear where the author intends to make claims about the "practical significance" (Kirk, 1996) of the results.

6. <u>Require effect size reporting</u>. Editors should require that effect size estimates be reported for all quantitative analyses. These are strongly *suggested* by APA (1994); however, Thompson (1996, p. 29--emphasis in original) suggests that other professional associations

that publish professional journals "venture beyond APA, and *require* such reports in all quantitative analyses."

7. Encourage or require replicability and "what if" analyses. As previously discussed, replicability analyses provide reasonable evidence to support (or disconfirm) the generalizability of the findings, something that SSTs do NOT do (Shaver, 1993; Thompson, 1994). "What if" analyses, if used regularly, will build in readers and authors a sense of always considering the sample size when conducting SSTs, and thereby considering the problems inherent, particularly to rather larger and rather small samples.

8. Require authors to avoid using SSTs where they are not appropriate. For example, as previously noted, *EPM* does not allow manuscripts to be published if SSTs accompany validity or reliability coefficients.

9. Encourage or require that power analyses or replicability analyses accompany statistically nonsignificant results. These analyses allow for the researcher to address power considerations or to determine if a result with a small sample has evidence of stability in cases in which and SST indicates a statistically nonsignificant result.

10. Utilize careful substantive and copyediting procedures. Careful copyediting procedures will serve to assure that very little sloppy language relative to SSTs will

end up in published manuscript. In addition to the suggestions mentioned above, editors will want to make sure language such as "highly significant" and "approaching significance" is edited out of the final copies of accepted manuscripts.

References

American Psychological Association. (1994). Publication manual of the American Psychological Association (4th ed.). Washington, Author.

Bakan, D. (1966). The test of significance in psychological research. Psychological Bulletin, 66, 423-437.

Berkson, J. (1942). Tests of significance considered as evidence. Journal of the American Statistical Association, 37, 325-335.

Campbell, J. P. (1982). Editorial: Some remarks from the outgoing editor. Journal of Applied Psychology, 67, 691-700.

Carver, R. P. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Carver, R. P. (1993). The case against statistical significance testing, revisited. Journal of Experimental Education, 61, 287-292.

Chow, S. L. (1988). Significance test or effect size? Psychological Bulletin, 70, 426-443.

Cohen, J. (1988). Statistical power analysis of the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1994). The earth is round (p < .05). American Psychologist, 49, 997-1003.

Craig, J. R., Eison, C. L., & Metze, L. P. (1976). Significance tests and their interpretation: An example utilizing published research and $\hat{\omega}^2$. Bulletin of the Psychonomic Society, 7, 280-282.

Cronbach, L. (1975). Beyond the two disciplines of scientific psychology. American Psychologist, 12, 671-684.

Daniel, L. G. (1988). [Review of Conducting educational research (3rd ed.)]. Educational and Psychological Measurement, 48, 848-851.

Daniel, L. G. (1989b, January). Use of the jackknife statistic to establish the external validity of discriminant analysis results. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston. (ERIC Document Reproduction Service No. ED 305 382)

Daniel, L. G. (1990). [Review of Basic statistical analysis (3rd ed.)]. Educational and Psychological Measurement, 50, 710-716.

Daniel, L. G. (1992a, April). Bootstrap methods in the principal components case. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED 346 135)

Daniel, L. G. (1992b, November). Perceptions of the quality of educational research throughout the twentieth century: A comprehensive review of the literature. Paper presented at the annual meeting of the Mid-South Educational Research Association, Knoxville, TN.

Daniel, L. G. (1997). Kerlinger's research myths: An overview with implications for educational researchers. Journal of Experimental Education, 65, 101-118.

Daniel, L. G., & Witta, E. L. (1997, March). Implications for teaching graduate students correct terminology for discussing validity and reliability based on a content analysis of three social science measurement journals. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Edwards, W. (1965). Tactical note on the relation between scientific and statistical hypotheses. Psychological Bulletin, 63, 400-402.

Ferrell, C. M. (1992, February). Statistical significance, sample splitting and generalizability of results. Paper presented at the annual meeting of the Southwest Educational Research Association. (ERIC Document Reproduction Service No.)

Gold, D. (1969). Statistical tests and substantive significance. American Sociologist, 4, 42-46.

Henkel, C. G. (1976). Tests of significance. Newbury Park, CA: Sage.

Holmes, C. B. (1990). The honest truth about lying with statistics. Springfield, IL: Charles C. Thomas.

Huberty, C. J. (1987). On statistical testing. Educational Researcher, 16(8), 4-9.

Huberty, C. J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. Journal of Experimental Education, 61, 317-333.

Huck, S. W., & Cormier, W. G. (1996). Reading statistics and research (2nd ed.). New York: HarperCollins.

Kerlinger, F. N. (1979). Behavioral research: A conceptual approach. New York: Holt, Rinehart and Winston.

Kerlinger, F. N. (1986). Foundations of behavioral research (3rd ed.). Fort Worth, TX: Holt, Rinehart and Winston.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 5, 746-759.

Kish, L. (1959). Some statistical problems in research design. American Sociological Review, 24, 328-338.

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. American Psychologist, 43, 635-642.

Mahoney, M. J. (1976). Scientist as subject: The psychological imperative. Cambridge, MA: Ballinger.

McDonald, R. (1985). Factor analysis and related methods. Hillsdale, NJ: Erlbaum.

Measurement and Evaluation in Counseling and Development. (1992). Guidelines for authors. Measurement and Evaluation in Counseling and Development, 25, 143.

Meehl, P. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.

Melton, A. (1962). Editorial. Journal of Experimental Psychology, 64, 553-557.

Mohr, L. B. (1990). Understanding significance testing. Newbury Park, CA: Sage.

Morrison, D. E., & Henkel, R. E. (1970). The significance test controversy--A reader. Chicago: Adeline.

Neuliep, J. W. (Ed.). (1991). Replication in the social sciences. Newbury Park, CA: Sage.

Pedhazur, E. J., & Schmelkin, L. P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Erlbaum.

Popham, W. J. (1993). Educational evaluation (3rd ed.). Boston, MA: Allyn and Bacon.

Rozeboom, W. M. (1960). The fallacy of the null-hypothesis significance test. Psychological Bulletin, 57, 416-428.

Schafer, W. D. (1990). Interpreting statistical significance. Measurement and Evaluation in Counseling and Development, 23, 98-99.

Schafer, W. D. (1991). Power analysis in interpreting statistical nonsignificance. Measurement and Evaluation in Counseling and Development, 23, 2-3.

Schafer, W. D. (1993). Interpreting statistical significance and nonsignificance, Journal of Experimental Education, 61, 383-387.

Shaver, J. (1985). Chance and nonsense. Phi Delta Kappan, 67(1), 57-60.

Shaver, J. (1993). What statistical significance testing is, and what it is not. Journal of Experimental Education, 61, 293-316

Shea, C. (1996). Psychologists debate accuracy of "significance" test. Chronicle of Higher Education, 42(9), A12,A19.

Shulman, L. S. (1970). Reconstruction of educational research. Review of Educational Research, 40, 371-393.

Thompson, B. (1989a). Asking "what if" questions about significance tests. Measurement and Evaluation in Counseling and Development, 22, 66-67.

Thompson, B. (1989b). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues.

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. Journal of Counseling and Development, 70, 434-438.

Thompson, B. (1993a). Foreword. Journal of Experimental Education, 61, 285-286.

Thompson, B. (1993b). The use of statistical significance tests in research: Bootstrap and other alternatives. Journal of Experimental Education, 61, 361-377.

Thompson, B. (1994). Guidelines for authors. Educational and Psychological Measurement, 54, 837-847.

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. Educational and Psychological Measurement, 55, 525-534.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25(2), 26-30.

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. Educational and Psychological Measurement, 56, 197-208.

Thompson, B., & Daniel, L. G. (1996). Seminal readings on reliability and validity: A "hit parade" bibliography. Educational and Psychological Measurement, 56, 741-745.

Tucker, M. L., & Daniel, L. G. (1992, January). Investigating result stability of canonical function equations with the jackknife technique. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, TX. (ERIC Document Reproduction Service No. ED 343 914)

Tyler, R. W. (1931). What is statistical significance? Educational Research Bulletin, 10, 115-118,142.

Winch, R. F., & Campbell, D. T. (1969). Proof? No. Evidence? Yes. The significance of tests of significance. American Sociologist, 4, 140-143.

Yates, F. (1951). The influence of Statistical Methods for Research Workers on the development of the science of statistics. Journal of the American Statistical Association, 46, 19-34.

APPENDIX

EPM MANUSCRIPT REVIEW FORM

epm.rev 7/6/95

Educational and Psychological Measurement

Manuscript Review Form

Reviewer Code #________ MS #________

Due Date: ____/____/____

Omit criteria that are not relevant in evaluating a given ms. Return the rating sheet and comments to the appropriate Editor in the attached return envelope.

Manuscripts under review should be treated as confidential, proprietary information (not to be cited, quoted, etc.). After review, the ms should be discarded.

Part I ("N.A." = Not Applicable) *Criteria associated with the editorials in the Winter, 1994 (vol. 54, no. 4), August, 1995 (vol. 55, no. 4), and April, 1996 (Vol. 56, no. 2) issues:*

[✓ YES]	NO	N.A.	For each reported statistical significance test, is an effect size also reported?
YES	[✓ NO]	N.A.	Is a null hypothesis test of no difference used to evaluate measurement statistics (e.g., concurrent validity or score reliability)?
[✓ YES]	NO	N.A.	If statistical significance tests are reported, were "what if" analyses of sample sizes presented?
YES	[✓ NO]	N.A.	In discussing score validity or reliability, do the au(s) ever use inappropriate language (e.g., "the test was reliable" or "the test was valid")?
[✓ YES]	NO	N.A.	If statistically non-significant results were reported, was either a power analysis or a replicability analysis reported?
YES	[✓ NO]	N.A.	Was a stepwise analysis conducted?

Part II *General Criteria*

Worst	1	2	3	4	5	Best	Noteworthiness of Problem
Worst	1	2	3	4	5	Best	Theoretical Framework
Worst	1	2	3	4	5	Best	Adequacy of Sample
Worst	1	2	3	4	5	Best	Appropriateness of Method
Worst	1	2	3	4	5	Best	Insightfullness of Discussion
Worst	1	2	3	4	5	Best	Interest to EPM readership
Worst	1	2	3	4	5	Best	Writing Quality

Part III *Overall recommendation* ("Full review" involves review of the revision by all initial referees)

_____ Accept "as is" or with very minor revisions

_____ Tentatively accept pending revisions reviewed by the editor

_____ Encourage major revision with full review of the revision

_____ Allow revision, require full review of the revision

_____ Reject

_____ Ms more appropriate for another journal:

Part IV *Please provide the au(s) with constructive suggestions, helpful references, and related comments*, attaching additional sheets as needed.

TM026584

AERA CONFERENCE 1997

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Statistical Significance Testing in Educational and Psychological Measurement and Other Journals

Author(s): Larry G. Daniel

Corporate Source: University of Southern Mississippi

Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following two options and sign at the bottom of the page.

☒

Check here
For Level 1 Release:
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical) *and* paper copy.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Larry G. Daniel
Sample
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 1

The sample sticker shown below will be affixed to all Level 2 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY
Sample
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 2

□

Check here
For Level 2 Release:
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical), but *not* in paper copy.

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Sign here→ please

Signature: Larry G. Daniel

Printed Name/Position/Title: Larry G. Daniel, Assoc. Professor

Organization/Address:
Educational Leadership + Research
Univ. of Southern Mississippi
Hattiesburg, MS 39406-5027

Telephone: 601 2665832

FAX: 601 2665141

E-Mail Address: larry-daniel@bull.cc.usm.edu

Date: 03/25/97

(over)